Ph.D. in Information Technology
Thesis Defense

February 23rd, 2024
at 14:30
Sala Seminari Nicola Schiavoni

Davide PIANTELLA – XXXVI Cycle

BIG DATA INTEGRATION: SUPPORTING VARIOUSLY-STRUCTURED DATA BY MEANS OF METADATA AND DATA LAKES

Supervisor: Prof.ssa **Letizia TANCA**

Abstract:

In the big data context, we have the possibility to access and analyze an incredible amount of data to take informed decisions, and this data may belong to different data sources, each of which with different quality. As a consequence, in order to ascertain that the data we are going to use are correct, we need to find a way to assess the trustworthiness of the data sources providing them.
Such an evaluation must surely be performed in the context of *data integration*, where we aim to align different datasets to provide uniform access to data, possibly addressing sources with different database schemata, heterogeneous data formats, semantic and representation ambiguity, and data inconsistency; indeed, when we have multiple data sources we may often observe *conflicting information* on the same topic, due to many different reasons such as outdated values, noisy data, ambiguities in the representation of the same information or, more trivially, errors.
We may need to apply the data integration process to structured data (i.e., datasets with a prescriptive, regular, and complete structure, like relational databases) but also to semi-structured and totally unstructured data, such as texts written in natural language. In fact, data sources provide more and more often values in textual form that, depending on their length, are very difficult or impossible to be integrated using the classical data-integration pipeline. In such difficult contexts, *metadata* (i.e., information describing different characteristics of the data itself) can be used to support the integration of data with high heterogeneity of formats.
An emerging trend used to store huge amounts of data is the *data lake* paradigm: one of the main features of data lakes is the possibility of storing raw data of different natures without any pre-processing, providing tools to combine and retrieve, on demand, relational data, texts, images, logs, streaming data, etc.
In this thesis we study how *metadata* and *data lakes* can be leveraged to address the complex problem of big data integration. Specifically, first we provide an algorithm computing and exploiting *source authority* metadata to assess the trustworthiness of data sources in the *data fusion* phase of the data integration process, where we resolve the value conflicts that arise when sources provide conflicting values for the same data item. Furthermore, we present different methods and techniques that can be used to ease the adoption of a data lake in the *healthcare* domain: a *minimum clinical metadataset*; a *pipeline to extract medical concepts* expressed in natural language texts; a new *clinical word embeddings* designed for the Italian language; and finally, a *synthetic healthcare data generator* to enhance the performance assessment of a data lake. The closing contribution of this thesis is a novel pipeline designed to address the integration of data sources containing dirty values composed of long natural language texts, leveraging sentence embeddings and clustering techniques.

PhD Committee

Prof. **Davide Martinenghi,** Politecnico di Milano

Prof**. Paolo Ciaccia,** Università di Bologna

Prof. **Tamer Ozsu**, University of Waterloo